Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 8, pp. 1595–1613
DOI : 10.47974/JSMS-1116

ANOVA F-test and Kruskal-Wallis test performance comparison under varying distributions, variance heterogeneity, sample sizes, and noncentrality structures

Ryan R. Rahrig
School of Science, Technology and Mathematics
Ohio Northern University
Ada, OH 45810
U.S.A.

Abstract

This study provides a direct comparison of the ANOVA F-test and the Kruskal-Wallis Test, a commonly used nonparametric alternative, in terms of type I error rate and power performance. The Monte Carlo simulation expands upon previous studies by varying all common relevant factors simultaneously. The study included normal and nonnormal distributions, homogeneous and heterogeneous variance structures, balanced and unbalanced sample size designs, and also four specific noncentrality structures. Results indicate that the Kruskal-Wallis test is optimal to its parametric counterpart in many specific types of configurations.

Subject Classification: *62G10 - Nonparametric Hypothesis Testing.*

Keywords: *ANOVA, F-test, Kruskal-Wallis test, Type I error, Power.*

1. Introduction

The ANOVA F-test is the most commonly chosen test for testing the equality of location parameters for two or more groups. It is well-known that the appropriateness of this test depends on several conditions: (i) independence of samples, (ii) population normality, and (iii) variance homogeneity. Studies have shown that violations to these assumptions appear to occur quite often [1–5]. Consequently, important emphasis has been placed on developing and evaluating various alternatives to the ANOVA F-test which do not rely on one or more of the conditions stated above [6–8]. Foremost among these, in terms of common usage, appears to

E-mail: `r-rahrig@onu.edu`

be the Kruskal-Wallis test [7]. In a survey of commonly used textbooks for introductory statistics, it is typically the only ANOVA alternative covered, if one is presented at all [9–16]. More advanced statistics textbooks tend to include the Kruskal-Wallis test along with one additional alternative, which differs from text to text [17–19].

Prior research has examined the relative performance of the ANOVA F-test and Kruskal-Wallis test [20–29]. However, these studies are limited in that they do not consider one or more of the following: (i) non-normally distributed populations [20,22], (ii) heterogeneous population variances [21,23,24,27], (iii) four groups [21,23–25,27], (iv) performance in terms of power [22,26]. Several also do not consider unbalanced designs [21–23,25] even though studies have shown that unbalanced designs are more common for one-way ANOVA experiments [30,31]. While none of the limitations stated above apply to [28,29], those studies utilize the exponential distribution as the only non-normal distribution and thus can be improved by conducting a study that involves a wider variety of distribution shapes. Several authors have noted in recent years that there is further need for research in this area [27,32,33].

This simulation study provides extensive analysis of the performance of these two tests in terms of power and type I error for varying degrees of non-normality, heterogeneity of variances, and noncentrality structures for the alternative hypotheses in the case of four groups. Balanced and unbalanced sample designs are included.

The aim of this paper is to expand upon previous studies by simultaneously examining performance of the two procedures under all possible combinations of five population distributions of varying characteristics, five variance structures, four sample size designs, and four noncentrality structures.

In the present study, a simulation experiment was performed to expand upon those previously conducted that would enable analysis of the simultaneous effects of nonnormality, sample size design, variance heterogeneity, and noncentrality structure on the performance of the two tests in terms of type I error rate and power in the case of four groups. All possible combinations of five population distributions of varying characteristics, five variance structures, four sample size designs, and four mean structures were considered.

2. Methodology

The specific levels of each design factor used in the simulation study will be described in this section following a brief discussion regarding the nature of the hypotheses considered by each test.

The ANOVA F-test is used to test the null hypothesis stating that two or more population means are all equal versus the alternative hypothesis claiming a difference in at least one of the population means. Although generally agreed upon to be a test of location, the forms of the specific hypotheses tested by the Kruskal-Wallis test are not as well-defined, and current statistical textbooks are inconsistent about what aspects of the populations can really be compared. Some authors consider the null hypothesis to be that all population distributions are identical [9,34,35]. Others state the null as equality of population medians [13,18,19]. Others present it simply as a test of location without getting into the specifics of the hypotheses [10,17], and still others present it as a test of population mean equality [12,19].

While the two procedures may not test precisely the same hypotheses in theory, the important point is that the Kruskal-Wallis test is typically presented as a nonparametric analogue to the ANOVA F-test and therefore is operationally used to test the same hypotheses in practice. For that reason, this simulation study will analyze the performance of the two procedures when being utilized as a test of equality for population means.

A wide variety of population shapes were chosen for this study using five distributions employed by Blair and Higgins in their comparative study of the Wilcoxon Rank-Sum test and Student's t-test [36]. The normal distribution was included as the first population for validation of results, to allow for comparison with prior studies that examined performance under this distribution, and to serve as a benchmark when analyzing performance with nonnormality.

The uniform distribution was selected as the second population distribution in order to represent one extreme in the family of symmetric power distributions while the other extreme was represented by the double exponential (Laplace) distribution, the third distribution selected. The fourth population included was the truncated (half) normal distribution since it is a nonsymmetric distribution whose tail matches that of a normal curve. The final population studied was the exponential distribution as an example of a nonsymmetric, skewed population whose tails differs from that of a normal curve.

All population means were set to zero for the study of type I error rates. For the study of power, the four configurations of population mean differences used in [20] were utilized for the noncentrality structures. The first of these four is the equally spaced means pattern (ES pattern) that occurs when there is a linear trend among the means (e.g., $\mu_1 = 0$, $\mu_2 = c$, $\mu_3 = 2c$, and $\mu_4 = 3c$). The second and third represented the situation of having one extreme mean, with all others being equal (EX pattern). In EX1, the first variant of the EX pattern, the first mean was the extreme mean (e.g., $\mu_1 = c$, $\mu_2 = 0$, $\mu_3 = 0$, and $\mu_4 = 0$). The second variant (EX2) assigned the extreme mean to the fourth group (e.g., $\mu_1 = c$, $\mu_2 = c$, $\mu_3 = c$, and $\mu_4 = 0$). The fourth is a configuration with two equal means halfway between two extreme means (2M pattern; e.g., $\mu_1 = -c$, $\mu_2 = 0$, $\mu_3 = 0$, and $\mu_4 = c$). Specific values of c were found empirically such that the power for the ANOVA F-test would be about .86, a value chosen in order to be consistent with what was done in prior studies of this nature [28].

To examine how the tests react to sample size, the four balanced and unbalanced sample size configurations employed in [20] were utilized. The four configurations consisting of small to moderately large sample sizes for n_1:n_2:n_3:n_4 were 20:20:20:20, 12:12:12:12, 30:24:16:10, and 18:14:10:6. Also, to be consistent with prior research, the five variance configurations utilized in [20] were selected to allow for performance evaluation under both variance homogeneity and various levels of variance heterogeneity present. These five variance structures used for σ_1^2:σ_2^2:σ_3^2:σ_4^2 were 6:6:6:6, 12:6:4:1, 6:3:2:1, 1:4:6:12, and 1:2:3:6. As noted by Tomarken, 'maximum ratios of 12:1 and 6:1 were chosen because they represent enough heterogeneity to encourage researchers to conduct heterogeneity of variance tests and/or consider alternatives to the ANOVA, but not so much heterogeneity as to be unrealistic.'

The Monte Carlo simulation study was implemented using R [37]. For each replication of the study, the R function *aov* was used to conduct the ANOVA F-test, and *kruskal.test* performed the Kruskal-Wallis test. Observations from the normal, uniform, double exponential, half normal, and exponential distributions were generated using the R functions *rnorm, runif, rdexp, rtnorm,* and *rexp,* respectively. The nimble package was required for *rdexp* [38], and the *msm* package was required for *rtnorm* [39]. All other functions come with the base installation of R.

Ten thousand replications of the experiment were performed at each of the combinations of design factors at significance level .05. Studies have shown that this number of replications is sufficient to ensure reliable results [40,41]. For each combination, the proportion of rejections for each test was computed as the empirical type I error rate (in the case of equal means) or the empirical power (in all other cases).

3. Results

3.1 *Type I Error Rate*

Tables 1-5 provide the simulation results for the assessment of type I error rates for the normal, uniform, double exponential, truncated normal, and exponential populations, respectively, under the variance homogeneity and heterogeneity cases in combination with the four sample size configurations. Concordance among the values in Table 1 with those of Table 2 in [20] serves as a validation of those results.

Table 1

Empirical Type I Error Rates Simulated Under the Normal Distribution

	$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
	6:6:6:6		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	F	KW
20:20:20:20	.054	.051	.070	.061	.060	.054				
12:12:12:12	.050	.047	.072	.062	.061	.050				
30:24:16:10	.045	.043	.026	.031	.025	.030	.139	.085	.161	.105
18:14:10:6	.051	.047	.027	.033	.027	.028	.139	.083	.165	.102

Table 2

Empirical Type I Error Rates Simulated Under the Uniform Distribution

	$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
	6:6:6:6		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	F	KW
20:20:20:20	.052	.050	.072	.068	.059	.056				
12:12:12:12	.051	.045	.068	.060	.065	.062				
30:24:16:10	.054	.050	.026	.031	.026	.028	.135	.097	.160	.113
18:14:10:6	.050	.046	.029	.030	.028	.029	.141	.097	.163	.111

Table 3

Empirical Type I Error Rates Simulated Under the Double Exponential Distribution

	$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 : \sigma_4^2$									
	6:6:6:6		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
n_1:n_2:n_3:n_4	**F**	**KW**	F	**KW**	**F**	**KW**	**F**	**KW**	F	KW
20:20:20:20	.047	.045	.067	.059	.055	.048				
12:12:12:12	.047	.044	.059	.055	.055	.052				
30:24:16:10	.049	.049	.023	.032	.026	.033	.131	.077	.152	.086
18:14:10:6	.052	.048	.024	.028	.023	.031	.130	.074	.145	.084

Table 4

Empirical Type I Error Rates Simulated Under the Truncated Normal Distribution

	$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 : \sigma_4^2$									
	6:6:6:6		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
20:20:20:20	.049	.049	.067	.116	.063	.092				
12:12:12:12	.047	.042	.073	.091	.069	.081				
30:24:16:10	.048	.049	.025	.056	.024	.054	.141	.133	.164	.173
18:14:10:6	.051	.045	.028	.045	.027	.044	.137	.112	.165	.140

	$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 : \sigma_4^2$									
	6:6:6:6		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
20:20:20:20	.046	.045	.071	.251	.065	.197				
12:12:12:12	.040	.043	.069	.156	.065	.132				
30:24:16:10	.045	.043	.030	.154	.030	.131	.136	.226	.158	.311
18:14:10:6	.046	.048	.036	.092	.031	.084	.136	.175	.151	.225

3.1.1. Homogeneous variances

The two tests performed quite similarly under variance homogeneity. In all 20 cases, the difference in empirical rejection rates was never more than .007 (all but two were .005 or under). In 17 of the 20 cases, the KW rate is lower than that of the F although these differences typically do not exceed the amount which can be attributed to sampling variability. In terms of robustness, both tests did quite well in that the robustness limit (rejection rate > .07) was never exceeded. In fact, the KW test never exceeded the nominal rate of .05, and the F test never exceeded .054.

3.1.2. Heterogeneous variances

For each of the five population shapes, results show general agreement within three classifications of heterogeneity/sample size configurations – positive pairings of unequal variances with unequal sample sizes yielded the lowest rejection rates, followed by configurations with equal sample sizes, and then negative pairings of unequal variances with unequal sample sizes. Table 6 illustrates the combinations of sample size and variance structures that fall under each of these classifications.

Table 6

Variance/Sample Size Configuration Groupings

<table>
<tr><td></td><td colspan="4">$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 : \sigma_4^2$</td></tr>
<tr><td>n_1:n_2:n_3:n_4</td><td>6:6:6:6</td><td>12:6:4:1 6:3:2:1</td><td>1:2:3:6</td><td>1:4:6:12</td></tr>
<tr><td>20:20:20:20</td><td rowspan="4">A</td><td rowspan="2">B</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>12:12:12:12</td></tr>
<tr><td>30:24:16:10</td><td rowspan="2">C</td><td colspan="2" rowspan="2">D</td></tr>
<tr><td>18:14:10:6</td></tr>
</table>

Note: A = homogeneous variance cases. B = balanced sample size heterogeneous variance cases. C = positively paired unbalanced sample size heterogeneous variance cases. D = negatively paired unbalanced sample size heterogeneous variance cases.

For the three symmetric distributions (Tables 1-3), the rates for KW and F are quite similar in both the positive pairing and equal sample size cases (maximum difference = .011). With balanced sample sizes, both tests had rates above the nominal level but none were above the robustness limit of .07. The KW rate was lower than that of the F in all of these cases. With positive pairings both tests were quite conservative, with KW being less so in all cases. Results were less favorable for the two methods when the variances and sample sizes were negatively paired. The empirical rejection rates exceeded the robustness limit in all cases with the F test producing rates ranging 0.130 and 0.165 while the KW test performed relatively better with rates ranging from 0.074 to 0.113.

For the two nonsymmetric distributions (Tables 4-5), the F-test achieved similar results to what was attained for the symmetric distributions in each of the respective configurations. However, the KW test performed worse in each configuration as compared to that for the symmetric distributions (much worse in several cases). In contrast to the symmetric distributions, KW had higher rates than F in all cases with the only exception being the negative pairing configuration under the truncated normal distribution (although both tests produced unacceptably high rates). The positive pairing combinations under the truncated normal distribution are the only cases in which the KW rate is below the robustness threshold. The F-test generally had rates below 0.07 for the balanced sample size and positive pairing combinations for both the truncated normal and exponential distributions.

3.2. *Power*

Tables 7-11 provide the simulation results for the assessment of power for the normal, uniform, double exponential, truncated normal, and exponential populations, respectively, for the four specific noncentrality structures under the variance homogeneity and heterogeneity cases in combination with the four sample size configurations. Recall that specific values within each of the mean structures were empirically selected that would produce a power rate near 0.86 for the F test in each case.

Table 7

Empirical Power Simulated Under the Normal Distribution

		$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
Mean		**6:6:6:6**		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
Structure	$n_1:n_2:n_3:n_4$	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
ES	20:20:20:20	.851	.828	.847	.869	.856	.858				
	12:12:12:12	.853	.824	.863	.876	.856	.857				
	30:24:16:10	.866	.841	.868	.921	.864	.897	.863*	.815*	.856*	.817*
	18:14:10:6	.855	.821	.864	.910	.853	.882	.868*	.816*	.853*	.805*

Contd...

EX1	20:20:20:20	.857	.830	.850	.746	.857	.754				
	12:12:12:12	.854	.817	.862	.750	.856	.747				
	30:24:16:10	.850	.828	.853	.807	.857	.808	.870*	.887*	.850*	.882*
	18:14:10:6	.853	.827	.854	.807	.851	.799	.852*	.870*	.859*	.885*
EX2	20:20:20:20	.851	.825	.854	.948	.865	.935				
	12:12:12:12	.852	.815	.856	.943	.852	.922				
	30:24:16:10	.858	.814	.866	.974	.854	.957	.855*	.671*	.850*	.665*
	18:14:10:6	.863	.784	.852	.952	.852	.927	.858*	.649*	.845*	.628*
2M	20:20:20:20	.855	.830	.855	.890	.856	.872				
	12:12:12:12	.846	.813	.851	.885	.859	.870				
	30:24:16:10	.859	.833	.854	.914	.848	.882	.853*	.833*	.862*	.850*
	18:14:10:6	.854	.818	.859	.901	.853	.874	.858*	.829*	.852*	.837*

Note: ES = equally spaced means. EX1 = first group mean is the extreme mean with the other three being equal. EX2 = fourth group mean is the extreme mean with the other three being equal. 2M = two equal means halfway between two extreme means (e.g., $\mu_1 = -c$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = c$). Values for means were used that would produce power for the ANOVA F-test near 0.86. *Due to the large type I error rates in these configurations, these power values are largely irrelevant and will not be discussed in the paper.

Table 8

Empirical Power Simulated Under the Uniform Distribution

		$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
Mean		**6:6:6:6**		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
Structure	n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
ES	20:20:20:20	.855	.790	.864	.703	.859	.730				
	12:12:12:12	.863	.787	.857	.727	.857	.753				
	30:24:16:10	.861	.795	.861	.760	.855	.770	.861*	.694*	.844*	.632*
	18:14:10:6	.856	.775	.845	.785	.844	.786	.858*	.713*	.871*	.685*
EX1	20:20:20:20	.859	.782	.851	.614	.852	.629				
	12:12:12:12	.856	.767	.854	.633	.856	.644				
	30:24:16:10	.854	.790	.854	.675	.856	.687	.856*	.791*	.857*	.747*
	18:14:10:6	.857	.781	.858	.700	.855	.709	.853*	.793*	.854*	.774*
EX2	20:20:20:20	.858	.784	.856	.834	.866	.850				
	12:12:12:12	.854	.763	.851	.853	.849	.844				
	30:24:16:10	.852	.745	.855	.896	.855	.874	.850*	.577*	.842*	.546*
	18:14:10:6	.852	.714	.857	.906	.855	.873	.848*	.579*	.847*	.555*
2M	20:20:20:20	.857	.791	.854	.714	.853	.740				
	12:12:12:12	.856	.782	.859	.747	.848	.756				
	30:24:16:10	.854	.784	.855	.754	.859	.767	.863*	.718*	.856*	.676*
	18:14:10:6	.859	.781	.850	.786	.861	.790	.850*	.724*	.861*	.703*

Table 9

Empirical Power Simulated Under the Double Exponential Distribution

		$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
Mean		**6:6:6:6**		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
Structure	n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
ES	20:20:20:20	.860	.938	.858	.976	.860	.964				
	12:12:12:12	.854	.912	.852	.958	.866	.952				
	30:24:16:10	.856	.930	.858	.985	.860	.975	.856*	.931*	.847*	.939*
	18:14:10:6	.856	.913	.869	.975	.870	.967	.861*	.913*	.857*	.924*
EX1	20:20:20:20	.867	.927	.858	.894	.853	.888				
	12:12:12:12	.856	.897	.858	.855	.849	.844				
	30:24:16:10	.854	.933	.850	.930	.851	.930	.867*	.966*	.851*	.971*
	18:14:10:6	.847	.913	.845	.902	.857	.908	.866*	.947*	.852*	.956*
EX2	20:20:20:20	.853	.922	.863	.994	.859	.983				
	12:12:12:12	.856	.894	.853	.981	.852	.969				
	30:24:16:10	.852	.879	.859	.996	.854	.985	.859*	.801*	.852*	.787*
	18:14:10:6	.855	.824	.852	.980	.869	.959	.848*	.709*	.855*	.708*
2M	20:20:20:20	.857	.930	.858	.979	.846	.962				
	12:12:12:12	.853	.909	.857	.962	.861	.951				
	30:24:16:10	.858	.929	.857	.982	.871	.977	.846*	.937*	.858*	.958*
	18:14:10:6	.854	.907	.853	.968	.856	.954	.855*	.923*	.860*	.940*

Table 10

Empirical Power Simulated Under the Truncated Normal Distribution

		$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
Mean		**6:6:6:6**		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
Structure	n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
ES	20:20:20:20	.861	.903	.845	.956	.861	.955				
	12:12:12:12	.846	.872	.853	.930	.856	.925				
	30:24:16:10	.852	.898	.867	.973	.851	.966	.862*	.708*	.859*	.593*
	18:14:10:6	.860	.884	.854	.945	.855	.938	.859*	.782*	.857*	.713*

Contd...

EX1	20:20:20:20	.855	.908	.852	.684	.857	.720				
	12:12:12:12	.857	.890	.848	.746	.841	.764				
	30:24:16:10	.858	.909	.851	.755	.854	.773	.858*	.951*	.857*	.956*
	18:14:10:6	.857	.887	.858	.821	.856	.832	.860*	.926*	.851*	.923*
EX2	20:20:20:20	.859	.857	.849	.902	.854	.903				
	12:12:12:12	.853	.829	.849	.943	.853	.913				
	30:24:16:10	.852	.807	.856	.970	.859	.939	.849*	.736*	.859*	.761*
	18:14:10:6	.854	.760	.861	.974	.854	.929	.851*	.679*	.854*	.683*
2M	20:20:20:20	.854	.896	.859	.721	.858	.775				
	12:12:12:12	.840	.866	.859	.839	.855	.836				
	30:24:16:10	.858	.891	.856	.832	.856	.831	.858*	.934*	.856*	.945*
	18:14:10:6	.858	.873	.856	.916	.869	.889	.857*	.895*	.855*	.907*

Table 11

Empirical Power Simulated Under the Exponential Distribution

		$\sigma_1^2:\sigma_2^2:\sigma_3^2:\sigma_4^2$									
Mean		**6:6:6:6**		**12:6:4:1**		**6:3:2:1**		**1:2:3:6**		**1:4:6:12**	
Structure	n_1:n_2:n_3:n_4	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**	**F**	**KW**
ES	20:20:20:20	.851	.980*	.852	.996*	.855	.995*				
	12:12:12:12	.849	.960*	.852	.983*	.856	.978*				
	30:24:16:10	.847	.982*	.861	.998*	.858	.997*	.856*	.883*	.862*	.739*
	18:14:10:6	.852	.965*	.852	.987*	.851	.985*	.865*	.924*	.855*	.889*
EX1	20:20:20:20	.851	.987*	.857	.887*	.855	.916*				
	12:12:12:12	.855	.968*	.858	.933*	.857	.937*				
	30:24:16:10	.857	.984*	.861	.938*	.859	.949*	.858*	.992*	.855*	.995*
	18:14:10:6	.849	.961*	.852	.958*	.852	.960*	.856*	.973*	.854*	.975*
EX2	20:20:20:20	.860	.945*	.847	.982*	.854	.974*				
	12:12:12:12	.851	.901*	.847	.981*	.854	.967*				
	30:24:16:10	.850	.880*	.851	.991*	.857	.976*	.857*	.853*	.844*	.871*
	18:14:10:6	.855	.815*	.855	.987*	.852	.948*	.854*	.761*	.848*	.782*
2M	20:20:20:20	.856	.980*	.853	.885*	.861	.922*				
	12:12:12:12	.852	.952*	.850	.955*	.854	.943*				
	30:24:16:10	.854	.973*	.863	.968*	.858	.954*	.856*	.989*	.850*	.991*
	18:14:10:6	.853	.954*	.853	.976*	.852	.960*	.860*	.969*	.858*	.972*

3.2.1. Homogeneous variances

As expected, the F test outperformed the KW test in all cases when the distribution was normal. It's widely known that the F test is the uniformly most powerful test when the variances are equal and the population is normal [42]. However, the power for KW tended to only be 0.02 to 0.04 lower in most of these cases and thus was competitive. That was not the case when the distribution was uniform though as the F test power was typically 0.06 to 0.08 higher in most cases.

As the rates for the double exponential distribution in Table 9 show, symmetry of the population distribution does not directly imply an advantage for the F test. The KW test outperformed the F test in 15 of the 16 cases, on average having 6% more power in these cases. When the truncated normal distribution was used (Table 10), the KW test achieved higher empirical power in all cases except for the 4 cases where the EX2 mean structure was used for the alternative hypothesis. The advantage of the KW test was greatest for the exponential distribution (Table 11). It again bested the F test in 15 of the 16 cases and attained an empirical power of more than 94% in 13 of those cases and more than 97% in 6 cases.

When analyzing the effect of the mean structures, the KW test performed similarly across three of the structures (ES, EX1, and 2M) with the EX2 structure resulting in the lowest power values. This is particularly true for the unbalanced sample size designs. For all five distributions, the lowest power values for the KW test were attained under this configuration. This suggests that the performance of the KW test is generally hindered when the group with the extreme population and small sample size coincide.

3.2.2. Heterogeneous variances

As with the type I error rates, when the variances were heterogeneous the power results were generally concordant within the balanced sample size, positively paired, and negatively paired configurations for each of the five distributions. However, since the type I error rates for the negatively paired configurations were all found to be unacceptably large (see Tables 1-5), neither the F nor KW test should be considered as a viable test in this situation. This in turn makes their respective powers largely irrelevant and for that reason will not be discussed. Similarly, the exponential distribution power results will not be considered due to the high type I error rates for the KW test in all cases (see Table 5).

Power assessments appear to be heavily influenced by both the mean structure for the alternative and also the population distribution. The specific sizes of the samples and variances within each of the sample and variance structures do not have as great of an impact. The forms of the structures are what appear to be more important.

The results in Table 7 show that population normality alone is not sufficient justification for using the F test when variance heterogeneity is present. Under normality, F test rejection rates were inferior than those of the KW test in all balanced and positive pairing cases except when the mean structure was EX1. In these cases the KW test typically had empirical power near or above 90%. However, in the EX1 configuration (the extreme mean was paired with the largest variance in these cases), the KW test had rates of only 80% or lower.

The KW test performed best relative to the F test when the distribution was double exponential (Table 9). It outperformed the F test in all balanced and positive pairing configuration for all four mean structures. In all cases except when the mean structure was EX1, KW attained a power of 95% or more.

At the other end, in terms of performance, the KW test performed most poorly when the distribution was uniform (Table 8). The KW rates of rejection were lower than the F test in all configurations except when the mean structure was EX2 with a positive pairing of the sample size and variance structures.

For the truncated normal distribution (Table 10), the KW type I error rates under the balanced sample size design were excessively high and thus the KW test should not be considered in this scenario. The positive pairing configuration is the only one thus considered. With the ES mean structure, the KW test performed better than the F test in all four cases with an average power of 96%. With the EX1 mean structure, the KW test performed worse in all four cases, with rates ranging from 75% to 83%. The KW test again did better under the EX2 structure having an average power of 95%. The 2M structure produced mixed results dependent on the sample sizes.

4. Summary

Table 12 provides a summary assessment of the results. For each combination of distribution type, mean structure, and variance/sample size grouping, the table provides which of the two tests, if any, was shown to be preferred according to the type I error and power results. If the power values were within 0.01, then neither was declared to be the preferred test and both tests were listed.

Table 12

Summary of Test Preferability for the Various Configurations

		Mean Structure			
Distribution	**Sample size/Variance Configuration**	**ES**	**EX1**	**EX2**	**E2M**
Normal	A: Homogeneous Variances	F (KW)	F (KW)	F	F (KW)
	B: Equal n, Heterogeneous Variances	F, KW	F	KW	KW (F)
	C: Unequal n, Positive Variance Pairings	KW	F	KW	F
	D: Unequal n, Negative Variance Pairings	---	---	---	---
Uniform	A: Homogeneous Variances	F	F	F	F
	B: Equal n, Heterogeneous Variances	F	F	F, KW	F
	C: Unequal n, Positive Variance Pairings	F	F	KW	F
	D: Unequal n, Negative Variance Pairings	---	---	---	---
Double Exponential	A: Homogeneous Variances	KW	KW	---	KW
	B: Equal n, Heterogeneous Variances	KW	KW*	KW	KW
	C: Unequal n, Positive Variance Pairings	KW	KW	KW	KW
	D: Unequal n, Negative Variance Pairings	---	---	---	---
Truncated Normal	A: Homogeneous Variances	KW	KW	F	KW*
	B: Equal n, Heterogeneous Variances	KW	F	---	F
	C: Unequal n, Positive Variance Pairings	KW	F	KW	F, KW**
	D: Unequal n, Negative Variance Pairings	---	---	---	---
Exponential	A: Homogeneous Variances	KW	KW	KW	KW
	B: Equal n, Heterogeneous Variances	---	---	---	---
	C: Unequal n, Positive Variance Pairings	---	---	---	---
	D: Unequal n, Negative Variance Pairings	---	---	---	---

Note: Listing of preferred test (if either deemed acceptable) for each configuration based on simulated type I error and power results. An empirical type I error rate was required to be deemed acceptable. Both tests listed if empirical power difference less than 0.01.

* Empirical power difference greater than 0.01 but less than 0.03

** F-test preferred in larger sample size case, KW preferred in smaller sample size case.

As before, the variance and sample size configurations were classified as either homogeneous variances, balanced samples with heterogeneous variances, unbalanced samples positively paired with heterogeneous variances, or unbalanced samples negatively paired with heterogeneous variances (Table 6). Except in one case (noted in the table), assessments within each of these groupings did not differ. This suggests that actual sizes for the samples and variances do not play as critical of a role as does the form of the structure in conjunction with the distribution and mean structure types.

The aim of this investigation was to examine the simultaneous effects of distribution type, sample size structure, variance structure, and mean structure. The results summarized in the table strongly suggest that interaction among these effects exists, and one ought to consider this when relying on other studies which only consider a more limited scope of factors.

Indeed, in [20], it was acknowledged that future investigations considering the dual effects of nonnormality and variance heterogeneity may arrive at conclusions different from what that study showed. The present study certainly builds on that study (and others that only considered the normal distribution) as it clearly reveals that distribution type has a large impact in terms of test preference and performance. While results vary for the truncated normal and normal distributions, the Kruskal-Wallis test dominates when the population shape is double exponential, and the F-test dominates when the population shape is uniform. With both of the latter two distributions being symmetrical, that suggests that population symmetry alone is not sufficient justification for choosing one test over the other. When the population was exponential, the Kruskal-Wallis test was found to be the superior method when variances were homogeneous although neither test was adequate under variance heterogeneity with this distribution.

In [29], it was concluded that non-parametric tests are substantially affected by the inhomogeneity of variance. Table 12 shows that this is also true of the F-test and in many cases more so than the Kruskal-Wallis test. Several cases can be observed where F is preferable for a given distribution type and mean structure under variance homogeneity, but KW is preferable in a case of variance heterogeneity for the same distribution type and mean structure.

The specific structure of the means under the alternative hypothesis also plays a large role in performance. For example, when the distribution was truncated normal with unequal sample sizes positively paired with unequal variances, the F test was preferable for the EX1 structure, the KW test was preferable for the ES and 2M structures, and neither test was acceptable for the EX2 structure. Several rows within the table show a variation in test preference suggesting the structure of the means is important.

5. Conclusion

In conclusion, the results have shown that the Kruskal-Wallis test is an adequate test and superior to the F-test in many situations, and its inclusion in textbooks as a viable alternative to the classic test is at least partially justified. And although it is sometimes presented as a test of population medians that should not be used to test population means (see e.g., [29]), the present study shows that it does in fact perform quite well when testing populations means under many different scenarios.

Circumstances under which neither the F-test nor the Kruskal-Wallis test performs adequately were also observed. For example, poor performance was observed for both tests under the configurations where the variances and sample sizes were negatively paired. However, it could be argued that the problem here lies more in the design of the experiment rather than the choice of statistical test since the usual strategy is to sample more from the groups with greater variation and this execution is just the opposite. Yet the fact is that these configurations do arise in practice, and for this reason statistics textbook may want to place a greater emphasis on including F-test alternatives in addition to the Kruskal-Wallis test that perform well under these circumstances.

References

[1] T. Micceri, “The unicorn, the normal curve, and other improbable creatures,” *Psychological Bulletin*, vol. 105, pp. 156–166 (1989).

[2] M. J. Blanca, J. Arnau, D. López-Montiel, M. M. Marín, and A. L. Salvador, “Skewness and kurtosis in real data samples,” *Methodology,* vol. 9, pp. 78-84 (2013).

[3] C. R. Harvey and A. Siddique, “Conditional Skewness in Asset Pricing Tests,” *The Journal of Finance*, vol. 55, pp. 1263–1295 (2000).

[4] K. Kobayashi, “Analysis of quantitative data obtained from toxicity studies showing non-normal distribution,” *J. Toxicol. Sci.,* vol. 30, pp. 127–134 (2005).

[5] W. J. van der Linden, "A Lognormal Model for Response Times on Test Items," *Journal of Educational and Behavioral Statistics*, vol. 31, pp. 181–204 (2006).

[6] M. B. Brown and A. B. Forsythe, "372: The Anova and Multiple Comparisons for Data with Heterogeneous Variances," *Biometrics*, vol. 30, p. 719 (1974).

[7] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *JASA*, vol. 47, pp. 583–621 (1952).

[8] B. L. Welch, "On the Comparison of Several Mean Values: An Alternative Approach," *Biometrika*, vol. 38, p. 330 (1951).

[9] D. S. Moore, G. P. McCabe, and B. A. Craig, Introduction to the Practice of Statistics, 9th ed. New York: W.H. Freeman, Macmillan Learning (2014).

[10] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer, Mathematical Statistics with Applications, 7th ed. Belmont, CA: Thomson Brooks/Cole (2008).

[11] B. Baldi and D. S. Moore, The Practice of Statistics in the Life Sciences, 4th ed. New York, NY: W.H. Freeman, Macmillan Learning (2018).

[12] A. G. Bluman, Elementary Statistics: A Step by Step Approach, 10th ed. New York, NY: McGraw-Hill Education (2018).

[13] M. F. Triola and L. Iossi, Elementary Statistics, 13th ed. United States: Pearson (2018).

[14] D. T. Larose, Discovering Statistics, 3rd ed. New York: W.H. Freeman & Company, a Macmillan Education imprint (2016).

[15] M. Sullivan, Fundamentals of Statistics: Informed Decisions Using Data, 5th ed. Boston: Pearson (2018).

[16] C. Warren, K. Denley, and E. Atchley, Beginning Statistics, 2nd ed. Charleston, SC: Hawkes Learning Systems (2014).

[17] J. J. Higgins, An Introduction to Modern Nonparametric Statistics, Pacific Grove, CA: Brooks/Cole (2004).

[18] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 5th ed. Boca Raton: CRC Press (2011).

[19] P. Sprent and N. C. Smeeton, Applied Nonparametric Statistical Methods, 4th ed. Boca Raton: Chapman & Hall/CRC (2007).

[20] A. Tomarken and R. Serlin, "Comparison of ANOVA Alternatives Under Variance Heterogeneity and Specific Noncentrality Structures," *Psychological Bulletin*, vol. 99, pp. 90–99 (1986).

[21] A. Khan and G. D. Rayner, "Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem," *Journal of Applied Mathematics and Decision Sciences*, vol. 7, pp. 187–206 (2003).

[22] K. Moder, "Alternatives to F-Test in One Way ANOVA in Case of Heterogeneity of Variances (A Simulation Study)," *Psychological Test and Assessment Modeling*, vol. 52 (2010).

[23] T. V. Hecke, "Power Study of ANOVA Versus Kruskal-Wallis Test," *Journal of Statistics and Management Systems*, vol. 15, pp. 241–247 (2012).

[24] H. Buning, "Robust Analysis of Variance," *Journal of Applied Statistics*, vol. 24, pp. 319–332 (1997).

[25] P. A. Lachenbruch and P. J. Clements, "ANOVA, Kruskal-Wallis, Normal Scores and Unequal Variance," *Communications in Statistics - Theory and Methods*, vol. 20, pp. 107–126 (1991).

[26] Z. Yusof, S. Abdullah, and S. S. Syed-Yahaya, "Comparing the Performance of Modified F t Statistic with ANOVA and Kruskal Wallis Test," (2013).

[27] B. Lantz, "The Impact of Sample Non-Normality on ANOVA and Alternative Methods," *Br J Math Stat Psychol*, vol. 66, pp. 224–244 (2013).

[28] B. J. Feir-Walsh and L. E. Toothaker, "An Empirical Comparison of the ANOVA F-Test, Normal Scores Test and Kruskal-Wallis Test Under Violation of Assumptions," *Educational and Psychological Measurement*, vol. 34, pp. 789–799 (1974).

[29] H. J. Keselman, J. C. Rogan, and B. J. Feir-Walsh, "An Evaluation of Some Nonparametric and Parametric Tests for Location Equality," *British Journal of Mathematical and Statistical Psychology*, vol. 30, pp. 213–221 (1977).

[30] C. Golinski and R. A. Cribbie, "The Expanding Role of Quantitative Methodologists in Advancing Psychology," *Canadian Psychology/Psychologie canadienne,* vol. 50, pp. 83–90 (2009).

[31] H. J. Keselman, C. J. Huberty, L. M. Lix, R. L. O'Neill, B. A. Tobin, and R. A. Olejnik, "Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses," Rev. Educ. Res., vol. 68, pp. 350-386 (1998).

[32] M. J. Blanca, R. Alarcón, J. Arnau, M. J. Bono, and A. L. M. L. Liu, "Non-normal data: Is ANOVA still a valid option?" *Psicothema*, vol. 29, pp. 552–557 (2017).

[33] E. Schmider, M. Ziegler, E. Danay, S. Beyer, and C. L. F. D. P. Bühner, "Is it really robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption," *Methodology*, vol. 6, pp. 147–151 (2010).

[34] R. T. Hurlburt, Comprehending Behavioral Statistics, 6th ed. Dubuque, IA: Kendall Hunt Publishing Company (2017).

[35] D. C. Howell, Statistical Methods for Psychology, 8th ed. Belmont, CA: Wadsworth Cengage Learning (2013).

[36] R. C. Blair and J. J. Higgins, "A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions," *J. Educ. Behav. Stat.*, vol. 5, pp. 309–335 (1980).

[37] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria (2020). Available: https://www.R-project.org/.

[38] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik, "Programming with models: writing statistical algorithms for general model structures with NIMBLE," *J. Comput. Graph. Stat.*, vol. 26, no. 2, pp. 403-413 (2017).

[39] C. H. Jackson, "Multi-state models for panel data: The msm package for R," *J. Stat. Softw.*, vol. 38, pp. 1–29 (2011).

[40] R. Bendayan, J. Arnau, M. J. Blanca, and R. Bono, "Comparación de los procedimientos de Fleishman y Ramberg et al. para generar datos no normales en estudios de simulación," *Anales de Psicología*, vol. 30, no. 1, pp. 364-371 (2014).

[41] R. R. Robey and R. S. Barcikowski, "Type I error and the number of iterations in Monte Carlo studies of robustness," *Br. J. Math. Stat. Psychol.*, vol. 45, pp. 283–288 (1992).

[42] H. Scheffé, The Analysis of Variance, Wiley Classics Library ed. New York: Wiley-Interscience Publication (1999).

Received December, 2022